REVIEW & INTERPRETATION

Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding and Seed Bank Management

J. C. Reif, A. E. Melchinger,* and M. Frisch

ABSTRACT

A proper choice of a dissimilarity measure is important in surveys investigating genetic relationships among germplasm with molecular marker data. The objective of our study was to examine 10 dissimilarity coefficients widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. In particular, we (i) investigated the genetical and mathematical properties of these coefficients, (ii) examined consequences of these properties for different areas of application in plant breeding and seed banks, and (iii) determined relationships between these 10 coefficients. The genetical and mathematical concepts of the coefficients were described in detail. A Procrustes analysis of a published data set consisting of seven CIMMYT maize populations demonstrated close affinity between Euclidean, Rogers', modified Rogers', and Cavalli-Sforza and Edwards' distance on one hand, and Nei's standard and Reynolds dissimilarity on the other hand. Our investigations show that genetical and mathematical properties of dissimilarity measures are of crucial importance when choosing a genetic dissimilarity coefficient for analyzing molecular marker data. The presented results assist experimenters to extract the maximum amount of information from genetic data and, thus, facilitate the interpretation of findings from molecular marker studies on a theoretically sound basis.

Quantifying the degree of dissimilarity among genera, species, subspecies, populations, and elite breeding materials is of primary concern in population genetics and plant breeding. Before 1970, measures of genetic dissimilarity between taxonomic units relied on pedigree analysis and morphological, physiological, or cytological markers, as well as biometric analyses of quantitative and qualitative traits, heterosis, or the segregation variance in crosses (Melchinger, 1999). During the following two decades, isozymes have successfully been employed in numerous taxonomic and evolutionary studies (Hamrick and Godt, 1990), but their use in other applications was hampered by the small number of polymorphic markers available.

Molecular markers, such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs), have meanwhile replaced isozymes and are heavily used for (i) detection of genetic relationships among different germplasm in seed banks and breeding programs (cf., Brummer, 1999), (ii) prediction of heterosis (cf., Melchinger, 1999), (iii) search for promising heterotic groups for hybrid breeding (cf., Reif et al., 2003), (iv) identification of duplicates in seed banks (cf., van Treuren et al., 2001), (v) assessment of the level of genetic diversity present in germplasm pools and its flux across time (cf., Dubreuil and Charcosset, 1998; Labate et al., 2003), and (vi) identification of essentially derived varieties in plant variety protection (cf., Smith et al., 1991; Lombard et al., 2000).

In these various applications, a proper choice of a similarity s or dissimilarity coefficient $d = 1 - s$ (following the terminology of Gower, 1985) is important and depends on factors such as (i) the properties of the marker system employed, (ii) the genealogy of the germplasm, (iii) the operational taxonomic unit (OTU) (adopting the terminology of Sneath and Sokal, 1973) under consideration (e.g., lines, populations), (iv) the objectives of the study, and (v) necessary preconditions for subsequent multivariate analyses.

In a recent review, Mohammadi and Prasanna (2003) discussed the use of six coefficients d for the analysis of dichotomous molecular marker data, but ignored those coefficients based on allele frequencies, which are especially suitable for codominant marker data. Several authors (Goodman, 1972; Gower, 1985; Gower and Legendre, 1986) investigated the mathematical properties and relationships among various coefficients d. However, the above mentioned surveys disregarded coefficients, which are based on specific genetic models and, therefore, suitable for studies with seed bank or plant breeding materials.

To successfully conduct molecular marker surveys with plant breeding and seed bank materials, a thorough knowledge of genetical and mathematical properties of coefficients d is of crucial importance. Therefore, the objective of our study was to examine 10 coefficients d widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. In particular, we (i) investigated the genetical and mathematical properties of these coefficients, (ii) examined consequences of these properties for different areas of application in plant breeding and seed banks, and (iii) determined relationships between these 10 coefficients.

Nature of Molecular Marker Data

We suggest the term *allelic informative* if allele frequencies can be determined from the molecular marker

Institute of Plant Breeding, Seed Science, and Population Genetics, Univ. of Hohenheim, 70593 Stuttgart, Germany. Received 28 Jan. 2004. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 45:1–7 (2005).

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: AFLP, amplified fragment length polymorphism; OTU, operational taxonomic unit; SSR, simple sequence repeat.

Table 1. Dissimilarity coefficients d for allelic informative marker data. p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith locus in the two operational taxonomic units under consideration, n_i is the number of alleles at the ith locus, and m refers to the number of loci.

Variable	Dissimilarity coefficient		Range	Property	
				Distance	Euclidean
d_E	$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Euclidean	$0, \sqrt{2m}$	yes	yes
d_R	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Rogers (1972)	0, 1	yes	no
d_W	$\frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	Modified Rogers' (Wright, 1978; Goodman and Stuber, 1983)	0, 1	yes	yes
d_{CE}	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}})}$	Cavalli-Sforza and Edwards (1967)	0, 1	yes	yes
d_{RE}	$-\ln(1-\sum_{i=1}^{m}(a-b)/c)$ $a=1/2\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2$ $b=(1/(2(2n-1)))(2-\sum_{j=1}^{n_i}(p_{ij}^2+q_{ij}^2))$ $c=\sum_{i=1}^{m}(1-\sum_{j=1}p_{ij}q_{ij})$	Reynolds et al. (1983)	$0, \infty$	no	no
d_{N72}	$-\ln\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}q_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}^2\sum_{i=1}^{m}\sum_{j=1}^{n_i}q_{ij}^2}}$	Nei (1972)	$0, \infty$	no	no
d_{N83}	$\frac{1}{m}\sum_{i=1}^{m}(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}})$	Nei et al. (1983)	0, 1	no	no

data. Marker data are denoted as *allelic noninformative* if this is not feasible. For instance, SSR data of individual genotypes are allelic informative. The AFLP data are mostly allelic noninformative although Geerlings et al. (1999), Piepho and Koch (2000), and Jansen et al. (2001) described methods to estimate allele frequencies and, thus, score AFLP data as allelic informative in specific situations.

Provided that molecular marker data are allelic informative, the estimates of coefficients d between OTUs can be calculated from the difference in the allele frequencies (Table 1). For allelic noninformative molecular marker data, coefficients d based on absence or presence of observation of bands or signals must be applied (Table 2).

Distance and Euclidean Properties

Consider a set of elements M and a function d: $M \times M \rightarrow \mathbb{R}$, assigning a real number to each pair of elements in M. A dissimilarity d is called a distance or metric, if for each element $i, j, k \in M$ the following three properties hold true (Gower, 1985):

$$d(i, j) \geq 0 \text{ and } d(i, j) = 0 \text{ if and only if } i = j, \quad [1]$$

$$d_{ij} = d_{ji}, \quad [2]$$

$$d_{ik} \leq d_{ij} + d_{jk}. \quad [3]$$

Some simple but important properties follow from this definition. All elements of a distance matrix with respect to a set of OTUs S must be defined and positive or zero. The matrix is symmetric and the triangle inequality (Eq. [3]) holds true for all triplets ($i, j, k \in S$). The latter means that the length of any side of a triangle constructed with the three elements ($i, j, k \in$ S) is less than or equal to the sum of the lengths of the other two sides, with equality occurring only when the triangle degenerates to a line.

The coefficient d is Euclidean if n points $P_i \in \mathbb{R}^n$ exist such that the Euclidean distance between P_i and P_j is d_{ij} for all $i, j, \in M$ (Gower and Legendre, 1986). An illustration of the Euclidean property is given by these authors.

The Euclidean property is important because it is

Table 2. Similarity coefficients for allelic noninformative marker data, where v_{ij} refers to the bands in common between two operational taxonomic units (OTUs) i and j, w_{ij} is the number of bands present in i and absent in j, x_{ij} is the number of bands present in j and absent in i, and y_{ij} is the number of bands both absent in i and j.

Variable	Similarity coefficient		Range	Property			
				$1-s$		$\sqrt{1-s}$	
				Distance	Euclidean	Distance	Euclidean
s_{SM}	$\frac{v_{ij}+y_{ij}}{v_{ij}+w_{ij}+x_{ij}+y_{ij}}$	simple matching	0, 1	yes	no	yes	yes
s_J	$\frac{v_{ij}}{v_{ij}+w_{ij}+x_{ij}}$	Jaccard (1908)	0, 1	yes	no	yes	yes
s_D	$\frac{2v_{ij}}{2v_{ij}+w_{ij}+x_{ij}}$	Dice (1945)	0, 1	no	no	yes	yes

an explicit or implicit assumption of many multivariate analysis methods such as principal coordinate analysis, also known as classical multidimensional scaling, hierarchical cluster analysis, classification, hierarchical classification, and graph theory (Gower, 1985). However, if a coefficient d is not Euclidean, then there exists a constant b greater than some minimal value such that the matrix with the elements $(d_{ij} + b)$ is Euclidean (Cox and Cox, 2000). The problem of finding such a constant b has been referred to for many years, Messick and Abelson (1956) being an early reference. Thus, the Euclidean property is desirable but the main criteria for the choice of a coefficient d are its genetical properties. Both the Euclidean and genetical properties will be investigated for the coefficients d (Tables 1 and 2).

Genetic Dissimilarity Coefficients for Allelic Informative Marker Data

Euclidean Distance

The Euclidean distance is defined as:

$$d_E = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij} - q_{ij})^2}, \quad [4]$$

where p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith locus in the two OTUs under consideration, n_i is the number of alleles at the ith locus, and m refers to the number of loci. The d_E ranges from zero to $\sqrt{2m}$, the limits being assumed when the two OTUs have identical allele frequencies or are fixed for different alleles. Thus, an obvious disadvantage is that d_E values from different studies cannot be compared directly because d_E depends on the number of marker loci assayed.

The d_E is appropriate if allelic informative marker data are available and the relationships between OTUs (populations or individuals) are investigated in combination with multivariate methods that require dissimilarities possessing the Euclidean property.

Rogers' Distance

Rogers' distance (Rogers, 1972) is a modification of d_E and was developed assuming no knowledge about evolutionary forces diverging the OTUs under consideration:

$$d_R = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij} - q_{ij})^2}. \quad [5]$$

The d_R is the average d_E across all loci standardized with the factor $\sqrt{1/2}$ to restrict the values to the interval [0, 1]. It is one only if two OTUs are fixed for different alleles, but if one or both OTUs are not fixed and they have no alleles in common, d_R is not equal to one. d_R fulfills the distance properties (Nei et al., 1983), but it is not Euclidean. This follows from the identity $d_R = 1 - s_{SM}$ for homozygous inbred lines and the fact that $1 - s_{SM}$ is not Euclidean (Gower and Legendre, 1986).

Assuming that (i) F_1 was the cross between two homozygous inbred lines P1 and P2 and (ii) O was an inbred offspring derived from the F_1 cross, Melchinger et al. (1991) showed that d_R fulfilled following two genetical properties:

$$d_R(F_1, P1) = d_R(F_1, P2) = d_R(P1, P2)/2, \quad [6]$$

$$d_R(P1, O) + d_R(P2, O) = d_R(P1, P2). \quad [7]$$

The first property can be illustrated geometrically as three points F_1, P1, and P2 forming a line with F_1 lying in its center.

On the basis of these two properties, Melchinger et al. (1991) derived theoretical results that d_R estimates between two homozygous inbreds are linearly related to the coancestry coefficient (Malecot, 1948). Consequently, d_R is suitable for studying the relationship between the genetic dissimilarity of inbreds based on allelic informative marker data and the coefficient of coancestry (Malecot, 1948). This linear relationship is also desired in surveys (i) investigating the assembly and validation of core collections and the identification of duplicates in seed banks and (ii) uncovering pedigree relationships among OTUs as needed for the detection of essentially derived varieties in plant breeding.

Modified Rogers' Distance

Wright (1978, p. 91) and Goodman and Stuber (1983) modified d_R by assigning each allele one dimension in the modified Rogers' distance:

$$d_W = \frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij} - q_{ij})^2}. \quad [8]$$

Obviously, $d_W = \frac{1}{\sqrt{2m}}$ and as an Euclidean distance with values in [0, 1] it can be used for the same applications as recommended for d_E. Like d_R, d_W is not equal to one in the case of multiple alleles, even if the two OTUs have no allele in common.

Consider two populations, $\pi 1$ and $\pi 2$, in Hardy-Weinberg equilibrium and their hybrid population $\pi 1 \times \pi 2$. On the basis of results of Falconer and Mackay (1996), and assuming biallelism and absence of epistasis, Melchinger (1999) derived the following relationship between the mean of these populations:

$$\Delta H(\pi 1 \times \pi 2) = \mu_{\pi 1 \times \pi 2} - (\mu_{\pi 1} + \mu_{\pi 2})/2 = \sum_i y_i^2 \delta_i = \sum_i d_{Wi}^2(\pi 1, \pi 2)\delta_i, \quad [9]$$

where ΔH is the panmictic-midparent heterosis (Lamkey and Edwards, 1999), δ_i is the dominance effect at QTL i, and y_i is the difference in gene frequencies. Consequently, a linear relationship between ΔH and d_W^2 is expected under the above conditions. Therefore, d_W^2 is especially suitable in studies based on allelic informative marker data for examining (i) the prediction of heterosis with genetic dissimilarities or (ii) the establishment of heterotic groups. Furthermore, d_W can be used for the same applications as suggested for d_E, owing to its Euclidean property.

Cavalli-Sforza and Edwards' Chord Distance

Cavalli-Sforza and Edwards (1967) developed a genetic distance to analyze blood group allele frequencies in human populations. In this coefficient, an OTU with allele frequencies $p_1, p_2, ..., p_n$ is represented by the vector $(\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_n})$. Such a vector is always of unit length and, thus, the OTU is located on a surface of a hypersphere with a radius of one considering one locus. The distance between two OTUs is then directly proportional to the length of the chord connecting the points representing the OTUs. In particular, for two OTUs with no allele in common, d_{CE} is equal to one (Wright, 1978, p. 91). For multiple loci, the distances of all loci are combined by applying the Pythagorean theorem in many dimensions, so that the square of the distance between the OTUs is given by the sum of squared distances for each locus:

$$d_{CE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(1 - \sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right)}. \quad [10]$$

The d_{CE} ranges from zero to one even in the case of multiple alleles, which is an advantage over d_R and d_W. It can be shown that:

$$d_{CE} = \frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\sqrt{p_{ij}} - \sqrt{q_{ij}}\right)^2}. \quad [11]$$

Thus, d_{CE} is similar to d_W except that it uses the square root of the allelic frequencies as coordinates and is consequently an Euclidean distance. The d_{CE} was developed based on Kimura's (1954) model of selective drift by assuming that (i) the mutation rate is small and (ii) variation in selection pressure is rapid and haphazard (no constant direction in allele frequency changes). It seems doubtful that seed bank and plant breeding materials have evolved according to this model because selection pressure is rather directed than rapid and haphazard. However, if allelic informative marker data are available and one can assume the selective drift model, then d_{CE} is a proper coefficient to investigate phylogenetic relationships among populations. Because d_{CE} is Euclidean, it can be used for the same tasks as proposed for d_E.

Reynolds' Dissimilarity

Reynolds et al. (1983) used the coancestry coefficient θ (Malecot, 1948) as the basis for a measure of genetic dissimilarity for short-term evolution, when the divergence between populations with a common ancestral population may be regarded as being caused solely by drift:

$$d_{RE} = -\ln(1 - \theta), \quad [12]$$

where

$$\theta = \frac{\sum_{i=1}^{m}\left\{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij} - q_{ij})^2 - \frac{1}{2(2n-1)}\left[2 - \sum_{j=1}^{n_i}(p_{ij}^2 + q_{ij}^2)\right]\right\}}{\sum_{i=1}^{m}(1 - \sum_{j=1}p_{ij}q_{ij})} \quad [13]$$

For populations completely fixed at each locus (i.e., two homozygous inbred lines) θ is equal to one and d_{RE} is undefined. Thus, d_{RE} is neither a distance nor Euclidean. The d_{RE} was developed assuming that an ancestral population was split into several subpopulations of the same size, which subsequently diverged due to drift. In such a situation, d_{RE} is expected to increase linearly with the time since the populations diverged (Weir, 1996, p. 91), that is, $d_{RE} \approx t/2N$, where N is the subpopulation size and t the time measured in generations after divergence of the two populations. Thus, if mutation and selection can be neglected, and drift is the major evolutionary force, then d_{RE} is an appropriate dissimilarity coefficient for investigating the phylogenetic relationships among populations based on allelic informative marker data.

A recent application of d_{RE} was described by Labate et al. (2003), who examined relationships among U.S. maize landraces with SSR markers and assumed that an ancestral population split into several subpopulations diverging mainly due to drift. Mutation is known to have only small effects on genetic diversity compared with other forces and, thus, can safely be ignored in short-term evolution scenarios. However, neglecting selection as an evolutionary force in plant breeding or in seed bank populations seems questionable in most instances.

Nei's Standard Genetic Dissimilarity

In contrast to d_{CE} and d_{RE}, where it is assumed that populations diverged due to random genetic drift, Nei (1972) suggested a dissimilarity coefficient based on mutation and drift, often referred to as Nei's standard dissimilarity. This measure is intended to estimate the average number of codon substitutions per locus and was defined as:

$$d_{N72} = -\ln\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}q_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}^2\sum_{i=1}^{m}\sum_{j=1}^{n_i}q_{ij}^2}}. \quad [14]$$

Nei (1978) extended d_{N72} with a bias factor. If two OTUs differ in all alleles, d_{N72} is not defined, because it becomes $-\ln 0$. Thus, d_{N72} is neither a distance nor Euclidean. d_{N72} was developed based on the infinite-allele model (Kimura and Crow, 1964) assuming that an ancestral population split into various subpopulations, which diverged due to drift and mutation. If (i) the mutation-drift balance is maintained throughout the evolutionary process, (ii) selection is absent, and (iii) the dissimilarity is not very large, then $d_{N72} = 2vt$, where v is the mutation rate per locus and generation and t is the time measured in generations after divergence of the two populations (Nei et al., 1983). Under the above conditions, d_{N72} is suitable for investigating phylogenetic relationships among populations based on allelic informative marker data but otherwise, the same constrains apply as for d_{RE}.

Nei et al.'s (1983) Dissimilarity

Assuming the infinite allele model (Kimura and Crow, 1964), Nei et al. (1983) suggested in a simulation study a dissimilarity coefficient, which is quite efficient in recov-

ering the true evolutionary tree when it is reconstructed from allele frequency data (Nei and Kumar, 2000):

$$d_{N83} = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right), \quad [15]$$

which equals d_{CE}^2. Nevertheless, the result of the simulation study depends heavily on the underlying evolutionary model of the simulation scenario. The d_{N83} was not developed based on a specific genetic model and it is neither a distance nor Euclidean. Thus, application of d_{N83} in surveys for detecting phylogenetic relationships among populations seems questionable. For homozygous inbred lines, $d_{N83} = d_R$, and hence it could be used for the same applications as d_R.

Genetic Dissimilarity Coefficients for Allelic Noninformative Marker Data

With allelic noninformative marker data and two OTUs under consideration, one can form a 2 by 2 table with entries v_{ij} (number of bands in common between both OTUs), w_{ij} (number of bands present in the ith OTU and absent in the jth OTU), x_{ij} (number of bands absent in the ith OTU and present in the jth OTU), and y_{ij} (number of bands absent from both OTUs).

The simple matching coefficient is one of the oldest similarity coefficients (Sneath and Sokal, 1973):

$$s_{SM} = \frac{v_{ij} + y_{ij}}{v_{ij} + w_{ij} + x_{ij} + y_{ij}}. \quad [16]$$

For homozygous inbred lines, $d_{SM} = 1 - s_{SM} = d_R$, and therefore can be used for the same applications as suggested for d_R.

Jaccard (1908) suggested the similarity coefficient:

$$s_J = \frac{v_{ij}}{v_{ij} + w_{ij} + x_{ij}}. \quad [17]$$

The Dice coefficient (Dice, 1945) is defined as:

$$s_D = \frac{2v_{ij}}{2v_{ij} + w_{ij} + x_{ij}}. \quad [18]$$

The dissimilarity $d_D = 1 - s_D$ is also called the Nei–Li distance (Nei and Li, 1979) and is related to $d_J = 1 - s_J$ by a monotonic function.

In contrast to s_{SM}, both s_J and s_D do not involve negative matches (y_{ij}). For instance, if the probability of nonamplification of bands is high and absence of bands in both OTUs cannot be interpreted as a common characteristic, it is appropriate to apply coefficients s excluding negative matches (s_J and s_D).

In contrast to $1 - s$, $\sqrt{1 - s}$ is for all three presented coefficients a distance and Euclidean (Gower and Legendre, 1986). Thus, they could be used to examine relationships among OTUs based on allelic noninformative marker data in combination with multivariate methods, the explicit or implicit assumption of which is a dissimilarity coefficient with the Euclidean property (Gower, 1985).

Table 3. Residual sum of square values obtained by a Procrustes analysis with a published data set of seven CIMMYT maize populations (Reif et al., 2003) for seven dissimilarity coefficients based on differences in allele frequencies [Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W), and Cavalli-Sforza and Edwards' (d_{CE}) distance and Reynolds' (d_{RE}), Nei's (1972) (d_{N72}), and Nei et al.'s (1983) (d_{N83}) dissimilarity coefficient].

	d_E	d_R	d_W	d_{CE}	d_{RE}	d_{N72}
d_R	0.0014					
d_W	0.0000	0.0014				
d_{CE}	0.0038	0.0047	0.0038			
d_{RE}	0.0592	0.0636	0.0592	0.0787		
d_{N72}	0.0307	0.0336	0.0307	0.0474	0.0103	
d_{N83}	0.0209	0.0233	0.0209	0.0228	0.0281	0.0172

Relationships among Dissimilarity and Similarity Coefficients

If (i) band absence or presence can be interpreted as two alleles of one locus and (ii) the OTUs under consideration are homozygous inbreds, then the following relationships exist between the s and d coefficients:

$$d_R = d_W^2 = d_{N83} = d_{CE}^2 = \frac{1}{2m}d_E^2 = 1 - s_{SM}. \quad [19]$$

Gower (1975) proposed a method of comparing different multivariate analyses of the same data set, also known as Procrustes analysis (Cox and Cox, 2000). We used this approach to illustrate the differences between the dissimilarity coefficients based on allele frequency differences (Table 1).

The Procrustes analysis is based on the pairwise comparison between two sets of dissimilarities, d_{ij} and d_{ij}^* ($i, j = 1, 2, ..., n$), among the same sample of n OTUs. Rather than concentrating on the distances themselves, geometric points P_i ($i = 1, ..., n$) of the n OTUs are constructed to give rise to all the interdistances, d_{ij}. The

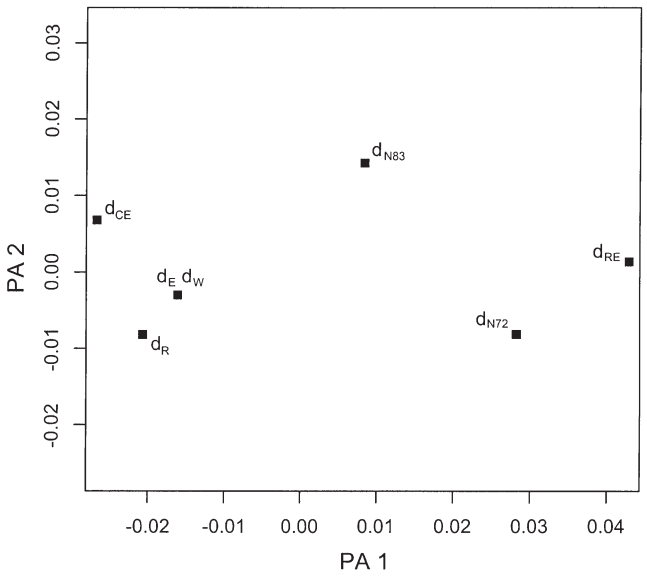

Fig. 1. First two principal axes (PA1 and PA2) of Kruskal's nonmetric multidimensional scaling for comparison of Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W) and Cavalli-Sforza and Edwards' distance (d_{CE}), and Nei's (1972) (d_{N72}), Nei et al.'s (1983) (d_{N83}), and Reynolds' (d_{RE}) dissimilarity coefficient based on a Procrustes analysis with a published data set of seven CIMMYT maize populations (Reif et al., 2003).

Table 4. Overview of the genetical and mathematical properties of dissimilarity coefficients based on allelic informative marker data: Euclidean (d_E), Rogers' (d_R), modified Rogers' (d_W) and Cavalli-Sforza and Edwards' distance (d_{CE}) and Nei's (1972) (d_{N72}), and Nei et al.'s (1983) (d_{N83}) and Reynolds' (d_{RE}) dissimilarity coefficient.

Dissimilarity coefficient	Properties
d_E	No underlying genetic concept. Suited to investigate relationships among operational taxonomic units (OTUs) with multivariate methods that require Euclidean distances (principal coordinate analysis, hierarchical cluster analysis, classification, hierachical classification, and graph theory).
d_R	Linearly related to coefficient of coancestry. Appropriate to examine (i) the assembly and validation of core collections and (ii) the uncovering of pedigree relationships among OTUs such as the detection of essentially derived varieties in plant breeding or the identification of duplicates and collection gaps in seed banks.
d_W	d_W^2 is linearly related to panmictic-midparent heterosis. Therefore, d_W^2 is appropriate to examine (i) the prediction of heterosis with genetic distances or (ii) the establishment of heterotic groups.
d_{CE}	Based on Kimura's (1954) model of selective drift. If one can assume the selective drift model, then d_{CE} is a proper coefficient to investigate the phylogenetic relationships among populations.
d_{RE}	Based on a model that an ancestral population splits into several subpopulations of the same size, which diverge due to drift. Thus, if mutation and selection can be neglected and drift is the major evolutionary force, then d_{RE} is suitable for investigating the phylogenetic relationships among populations.
d_{N72}	Based on the infinite-allele model (Kimura and Crow, 1964). If one can assume the infinite-allele model, then d_{N72} is suitable for investigating phylogenetic relationships among populations.
d_{N83}	For homozygous inbred lines, $d_{N83} = d_R$ and, hence, d_{N83} is also linearly related to the coancestry coefficient (Malecot, 1948). Therefore, d_{N83} can be used for inbred lines for the same applications as d_R.

coordinates of these points were obtained with Kruskal's nonmetric multidimensional scaling (Cox and Cox, 2000). Kruskal's nonmetric multidimensional scaling is a technique to represent OTUs in a reduced space while preserving the distance relationships among them with high fidelity. It is not limited to Euclidean distance matrices and can produce ordinations of objects from any dissimilarity matrix. Similarly, the coordinates of the points $P_i (i = 1, ..., n)$ are found for the dissimilarities d_{ij}^* by applying again Kruskal's nonmetric multidimensional scaling. The two configurations are then matched for best fit by means of translation, rotation, and reflection. The criterion of best fit adopted is the minimization of the residual sum of squares $R^2 = \sum_{i=1}^{2} d_E^2(P_i, P_i^*)$, where $d_E(P_i, P_i^*)$ is the Euclidean distance between corresponding points P_i and P_i^*.

We compared the seven coefficients d based on allele frequency differences (Table 1) of a published data set of seven tropical CIMMYT maize populations (Reif et al., 2003) by subjecting them pairwise to the Procrustes analysis. The resulting R^2 matrix (Table 3) was then used as input for Kruskal's nonmetric multidimensional scaling (Fig. 1). The same analyses were also performed with other data sets and yielded similar results (data not shown). All analyses were performed with Version 2 of the Plabsim software (Frisch et al., 2000), which is implemented as an extension to the statistical software R (Ihaka and Gentleman, 1996).

The distance between d_E and d_W is zero (Table 3) because $d_W = \sqrt{2m} d_E$. Both measures clustered together with d_R and d_{CE} (Fig. 1). This is in accordance with the expectations, because (i) d_{CE} equals d_W except that the square roots of the allele frequencies are used as coordinates and (ii) d_R is the average d_E across all loci standardized by the factor $\sqrt{2}$. The d_{N83} was positioned between d_E, d_W, d_R, and d_{CE} on one side and d_{N72} and d_{RE} on the other side. This is not surprising because d_{N72} and d_{RE} are based on similar assumptions: an ancestral population split into subpopulations diverging by drift (d_{RE}) or by mutation and drift (d_{N72}). Both coefficients include an estimate of the allele frequencies of the ancestral population in contrast to the other measures. Consequently, our results indicate that the analogy of d_{N72} and d_{RE} in estimating the allele frequencies of the ancestral population has a stronger influence on the property of the coefficients than the choice of the evolutionary model, assuming drift and mutation or only drift. Summarizing, some coefficients are mathematically related or were developed assuming similar evolutionary models.

CONCLUSIONS

Our investigations show that genetical (Table 4) and mathematical (Tables 1 and 2) properties of dissimilarity measures are of crucial importance when choosing a genetic dissimilarity coefficient for analyzing molecular marker data. The presented results can assist experimenters in the choice of dissimilarity measures that allow the extraction of the maximum amount of information from genetic data for given objectives. Thus, they facilitate the interpretation of findings from molecular marker studies on a theoretically sound basis.

REFERENCES

Brummer, E.C. 1999. Capturing heterosis in forage crop cultivar development. Crop Sci. 39:943–954.

Cavalli-Sforza, L.L., and A.W.F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19:233–257.

Cox, T.F., and M.A.A. Cox, 2000. Multidimensional scaling. v. 88. CRC Press, Boca Raton, FL.

Dice, L.R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297–302.

Dubreuil, P., and A. Charcosset. 1998. Genetic diversity within and among maize populations: A comparison between isozyme and nuclear RFLP loci. Theor. Appl. Genet. 96:577–587.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Group, London.

Frisch M., M. Bohn, A.E. Melchinger. 2000. Plabsim: Software for simulation of marker-assisted backcrossing. J. Hered. 91:86–87.

Geerlings, H., A.J. Van Oeveren, J.E. Pot, R.C. van Schaik. 1999. AFLP-Quantar Pro Image analysis software. http://www.keygene-products.com (verified 11 Aug. 2004). Keygene, Wageningen, the Netherlands.

Goodman, M.M. 1972. Distance analysis in biology. Syst. Zool. 174–186.

Goodman, M.M., and C.W. Stuber. 1983. Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187.

Gower, J.C. 1975. Generalised Procrustes analysis. Psychometrika 40:33–50.
Gower, J.C. 1985. Measures of similarity, dissimilarity and distances. p. 397–405. *In* S. Kotz, et al. (ed.) Encyclopedia of statistical sciences. Vol. 5. Wiley, New York.
Gower, J.C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3:5–48.
Hamrick, J.L., and M.J.W. Godt. 1990. Allozyme diversity in plant species. p. 43–63. *In* A.H.D. Brown, M.T. Clegg, A.L. Kahler, and B.S. Weir (ed.) Plant Population Genetics, Breeding and Genetic Resources. Sinauer, Sunderland, MA.
Ihaka, R., and R. Gentleman. 1996. A language for data analysis and graphics. J. Comput. Graphical Stat. 3:299–314.
Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44:223–270.
Jansen, R.C., H. Geerlings, A.J. van Oeveren, and R.C. van Schaik. 2001. A comment on codominant scoring of AFLP markers. Genetics 158:925–926.
Kimura, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics 39:280–295.
Kimura, M., and J.F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.
Labate, J.A., K.R. Lamkey, S.H. Mitchell, S. Kresovich, H. Sullivan, and J.S.C. Smith. 2003. Molecular and historical aspects of Corn Belt dent diversity. Crop Sci. 43:80–91.
Lamkey, K.R., and J.W. Edwards. 1999. Quantitative genetics of heterosis. Chapter 10. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.
Lombard, V., C.P. Baril, P. Dubreuil, F. Blouet, and D. Zhang. 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Sci. 40:1417–1425.
Malecot, G. 1948. Les Mathématiques de l'Hérédite. Masson et Cie, Paris.
Melchinger, A.E. 1999. Genetic diversity and heterosis. p. 99–118. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.
Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman, and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678.
Messick, S.M., and R.P. Abelson. 1956. The additive constant problem in multidimensional scaling. Psychometrica 21:1–15.
Mohammadi, S.A., and B.M. Prasanna. 2003. Analysis of genetic diversity in crop plants—Salient statistical tools and considerations. Crop Sci. 43:1235–1248.
Nei, M. 1972. Genetic distance between populations. Am. Nat. 106: 283–292.
Nei, M. 1978. The theory of genetic distance and evolution of human races. Jpn. J. Hum. Genet. 23:341–369.
Nei, M., and S. Kumar. Molecular evolution and phylogenetics. 2000. Oxford Univ. Press, New York.
Nei, M., and W.H. Li. 1979. Mathematical models for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269–5273.
Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. 19:153–170.
Piepho, H.-P., and G. Koch. 2000. Codominant analysis of banding data from a dominant marker system by normal mixtures. Genetics 155:1459–1468.
Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43:1275–1282.
Reynolds, J., B.S. Weir, C.C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. Genetics 105:767–779.
Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. p. 145–153. *In* Studies in genetics VII. Publ. 7213. Univ. of Texas, Austin.
Smith, J.S.C., O.S. Smith, S.L. Bowen, R.A. Tenborg, and S.J. Wall. 1991. The description and assessment of distances between inbred lines of maize. III. A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. Maydica 36:213–226.
Sneath, P.H.A., and R.R. Sokal. 1973. Numerical taxonomy. Freeman, San Francisco, CA.
van Treuren, R., L.J.M. van Soest, and Th.J.L. van Hintum. 2001. Marker-assisted rationalisation of genetic resources collections: A case study in flax using AFLPs. Theor. Appl. Genet. 103:144–152.
Weir, B.S. 1996. Genetic data analysis II. 2nd ed. Sinauer Assoc., Sunderland, MA.
Wright, S. 1978. Evolution and genetics of populations. Vol. IV. The Univ. of Chicago Press.